When to use agreement versus reliability measures

de Vet, H.C.W.; Terwee, C.B.; Knol, D.L.; Bouter, L.M.

published in
Journal of Clinical Epidemiology
2006

DOI (link to publisher)
10.1016/j.jclinepi.2005.10.015

document version
Publisher's PDF, also known as Version of record

ELSEVIER

Journal of Clinical Epidemiology 59 (2006) 1033–1039

Journal of Clinical Epidemiology

When to use agreement versus reliability measures

Henrica C.W. de Vet[a,*], Caroline B. Terwee[a], Dirk L. Knol[a,b], Lex M. Bouter[a]

[a]*Institute for Research in Extramural Medicine, VU University Medical Center, Amsterdam, Van der Boechorststraat 7, Amsterdam 1081 BT, The Netherlands*

[b]*Department of Clinical Epidemiology and Biostatistics, VU University Medical Center, Amsterdam, The Netherlands*

Accepted 25 October 2005

Abstract

Background: Reproducibility concerns the degree to which repeated measurements provide similar results. Agreement parameters assess how close the results of the repeated measurements are, by estimating the measurement error in repeated measurements. Reliability parameters assess whether study objects, often persons, can be distinguished from each other, despite measurement errors. In that case, the measurement error is related to the variability between persons. Consequently, reliability parameters are highly dependent on the heterogeneity of the study sample, while the agreement parameters, based on measurement error, are more a pure characteristic of the measurement instrument.

Methods and Results: Using an example of an interrater study, in which different physical therapists measure the range of motion of the arm in patients with shoulder complaints, the differences and relationships between reliability and agreement parameters for continuous variables are illustrated.

Conclusion: If the research question concerns the distinction of persons, reliability parameters are the most appropriate. But if the aim is to measure change in health status, which is often the case in clinical practice, parameters of agreement are preferred.

Keywords: Agreement; Measurement error; Measurement instruments; Reliability; Repeated measurements; Reproducibility

1. Introduction

Outcome measures in medical sciences may concern the assessment of radiographs and other imaging techniques, biopsy readings, the results of laboratory tests, the findings of physical examinations, or the scores on questionnaires collecting information, for example, on functional limitations, pain coping styles, and quality of life. An essential requirement of all outcome measures is that they are valid and reproducible or reliable [1,2].

Reproducibility concerns the degree to which repeated measurements in stable study objects, often persons, provide similar results. Repeated measurements may differ because of biologic variation in persons, because even stable characteristics often show small day-to-day differences, or follow a circadian rhythm. Other sources of variation may originate from the measurement instrument itself, or the circumstances under which the measurements take place. For instance, some instruments may be temperature dependent, or the mood of a respondent may influence the answers on a questionnaire. Measurements based on assessments made by clinicians may be influenced by intrarater or interrater variation.

This article first presents an example of an interrater study, then describes the concepts underlying various reproducibility parameters, which can be distinguished in reliability and agreement parameters. The primary aim of this article is to demonstrate the relationship and the important difference between parameters of reliability and agreement, and to provide recommendations for their use in medical sciences.

2. An example

In an interrater study on the range of motion of a painful shoulder, different reproducibility parameters were used to present the results [3]. To assess the limitations in passive glenohumeral abduction movement, the range of motion of the arm was measured with a digital inclinometer, and

* Corresponding author. Tel.: +31 20 444 8176; fax: +31 20 444 6775.

E-mail address: hcw.devet@vumc.nl (H.C.W. de Vet).

0895-4356/06/$ – see front matter
doi: 10.1016/j.jclinepi.2005.10.015

expressed in degrees. Two physical therapists (PT_A and PT_B) measured the range of motion of the affected and the nonaffected shoulder in 155 patients with shoulder complaints. Table 1 presents the results in terms of means and standard deviations, percentages of agreement within 5° and 10°, limits of agreement, and intraclass correlation coefficients (ICC) [3].

The first two lines in Table 1 present the means and standard deviations of the scores assessed by PT_A and PT_B for the affected and nonaffected shoulder. The standard deviations (SD) show the variability in the results between the patients, in which the heterogeneity of the study sample is reflected with regard to the characteristic under study. The third line presents the mean differences ($Mean_{diff}$) between the scores of PT_A and PT_B, and the SDs of these differences (SD_{diff}). The fourth and fifth lines present agreement parameters by reporting the percentages of patients for whom the scores of PT_A and PT_B differed less than 5° and 10°, respectively. For 43% of the patients the scores of PT_A and PT_B were within the 5° range, and for 72% they were within the 10° range, for both the affected and the nonaffected shoulder. The limits of agreement, calculated according to the Bland and Altman method [4], are also about similar for both shoulders and were −18.80° to 20.40° for the affected shoulder and −17.88° to 19.68° for the nonaffected shoulder. The last line shows the ICC, which is a parameter of reliability. There is quite a difference in the value of the ICC for the two shoulders: 0.83 for the affected shoulder and 0.28 for the nonaffected shoulder. For interpretation, the physical therapists would be quite satisfied with an agreement percentage of 72% of the patients within the 10° range, while the ICC (>0.7 is generally considered as good [5]) shows a satisfactory value for the affected shoulder, but not for the nonaffected shoulder. The explanation for these apparently contradictory results can be found in the conceptual difference between the two types of parameters.

Table 1
Reproducibility of the measurement of glenohumeral abduction of the shoulder

Parameters	Affected shoulder	Nonaffected shoulder
PT_A: Mean (SD)	69.49° (17.60°)	79.78° (7.60°)
PT_B: Mean (SD)	68.69° (16.25°)	78.88° (8.38°)
$Mean_{diff\_AB}$ (SD_{diff_AB})	0.80° (10.00°)	0.90° (9.58°)
PT_A vs. PT_B: % within 5°	43%	43%
PT_A vs. PT_B: % within 10°	72%	72%
Limits of $agreement_{\_AB}$	−18.80°–20.40°	−17.88°–19.68°
$ICC_{agreement\_AB}$	0.83	0.28

Abbreviations: ICC: intraclass correlation coefficient; $Mean_{diff\_AB}$: Mean of the differences between PT_A and PT_B; PT_A: physical therapist A; PT_B: physical therapist B; SD: standard deviation; SD_{diff_AB}: standard deviation of the differences between PT_A and PT_B.

3. Conceptual difference between agreement and reliability parameters

In the literature, agreement and reliability parameters are often used interchangeably, although some authors have pointed out the differences [6,7].

Agreement and reliability parameters focus on two different questions:

1. ‘‘How good is the agreement between repeated measurements?’’ This concerns the measurement error, and assesses exactly how close the scores for repeated measurements are.
2. ‘‘How reliable is the measurement?’’ In other words, how well can patients be distinguished from each other, despite measurement errors. In this case, the measurement error is related to the variability between study objects.

As an umbrella term for the concepts of agreement and reliability we use the term ‘‘reproducibility’’ [7], because both concepts concern the question of whether measurement results are reproducible in test–retest situations. The repetitions may concern different measurement moments, different conditions, different raters, or the same rater at different times.

Figure 1 visualizes the distinction between agreement and reliability. The weight of three persons is measured on 5 different days. The five measurements per person show some variation. The SD of the values of the repeated measurements of one person represents the agreement, and answers question 1 above. For reliability this measurement error is related to the variability between persons, and tells us how well they can be distinguished from each other. If the values of persons are distant (as for persons ● and ■), the measurement error will not affect discrimination of the persons, but if the values of persons are close (as for persons ■ and ▲) the measurement error will affect the ability to discriminate and the reliability will be substantially lower.

A reliability parameter (e.g., the ICC) has as typical basic formula:

$$\text{reliability} = \frac{\text{variability between study objects}}{\text{variability between study objects} + \text{measurement error}}$$

The reliability parameter relates the measurement error to the variability between study objects, in our case persons.

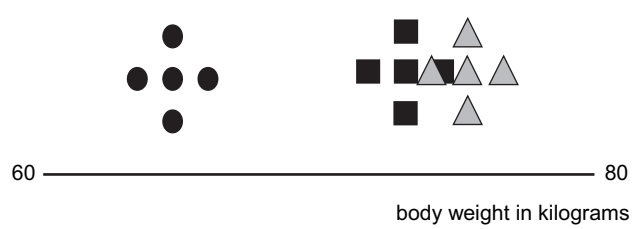

Fig. 1. Five repeated measurements of the body weights of three persons (●, ■, and ▲).

If the measurement error is small, compared to the variability between persons, the reliability parameter approaches 1. This means that the discrimination of the persons is hardly affected by measurement error, and thus the reliability is high (persons ● and ■ in Fig. 1). If the measurement error is large, compared to the variability between persons, the ICC value becomes smaller. For example, if the measurement error equals the variability between persons, the ICC becomes 0.5. This means that the discrimination will be affected by the measurement error (e.g., persons ■ and ▲ in Fig. 1). The ICC is a ratio ranging in value between 0 (representing a totally unreliable measurement) and 1 (implying perfect reliability).

We now turn back to our example. The percentage of scores of both PTs within 5° and 10° is a parameter of agreement: it estimates the measurement error. Note that it remains unclear whether this measurement error originates from the PTs, for example, because of a difference in the way they use the inclinometer, from variation within the patients who differ in their range of motion at the two moments of measurements, or from a combination of PTs and patients, for instance because of variation in the way the patients are motivated by the PTs. However, this measurement error is totally independent of the variability between persons. In our example, the agreement of the scores of the two PTs is approximately the same for the affected and the nonaffected shoulder. The ICC, which is a reliability parameter, relates this measurement error to the variability between persons in the population sample under study. The higher value of the ICC for the affected shoulder is explained by the higher variability between persons in glenohumeral abduction of the affected shoulder, compared to the nonaffected shoulder. This can be seen from the much larger standard deviation in the measurements of the affected shoulder, compared to the standard deviation for the nonaffected shoulder (first two lines in Table 1): patients all have maximum movement ability in the arm on the nonaffected side, but they differ considerably in the range of motion of the arm on the affected side. As the variability in scores for the affected shoulders is greater, these can more easily be distinguished, despite the same magnitude of measurement error. This clearly illustrates the difference between agreement and reliability parameters.

4. Agreement parameters are neglected in medical sciences

In the 1980s Guyatt et al. [8] clearly emphasized the distinction between reliability and agreement parameters. They explained that reliability parameters are required for instruments that are used for discriminative purposes and agreement parameters are required for those that are used for evaluative purposes. With a hypothetic example they eloquently demonstrated that discriminative instruments require a high level of reliability: that is, the measurement error should be small in comparison to the variability between the persons that the instrument needs to distinguish. Thus, if the differences between persons are large, a certain amount of measurement error is acceptable. For an evaluative measurement instrument the variability between persons in the population sample does not matter at all; only the measurement error is important. This measurement error should be smaller than the improvements or deteriorations that one wants to detect. If the measurement error is large, then small changes cannot be distinguished from measurement error. The smaller the measurement error, the smaller the changes that can be detected beyond measurement error.

In medical sciences measurement instruments are often used to evaluate changes over time, either with or without interventions. Nevertheless, many researchers still prefer reliability parameters over agreement parameters. In two recent clinimetric reviews [9,10] we assessed the quality of measurement instruments in terms of reproducibility. We registered whether agreement and reliability parameters were assessed. The measurement instruments were questionnaires to assess shoulder disability [9] or quality of life (QoL) in visually impaired persons [10]. These instruments are mainly used to evaluate the effects of interventions or monitor changes over time. Thus, these are typically evaluative measurement instruments. In the review of QoL of visually impaired persons [10] we found 31 questionnaires. For 16 questionnaires a reliability parameter was reported, but a parameter for agreement was presented for only seven questionnaires. For all 16 shoulder disability questionnaires a parameter of reliability was presented, but an additional parameter of agreement was presented for only for six questionnaires [9]. Apparently, agreement parameters have not yet struck root in medical sciences.

Streiner and Norman [1] argue that there is no need for a special parameter for measurement error, because it can be derived from the ICC formula. However, this is only true if all the components of the ICC formula are presented. Usually only the ICC value is provided, without even mentioning which ICC formula has been used [11], that is, with or without inclusion of the systematic difference between measurements. Even more important, authors who present only reliability parameters and no parameters of agreement usually draw the wrong conclusions: they rely solely on reliability parameters when they should have relied on parameters of agreement when evaluation is at issue.

5. Relationship between the agreement and reliability parameters

The relationship between parameters of agreement and reliability can best be illustrated by elaborating on the variances that are involved in the ICC formulas. Therefore, we first need to explain the meaning of the variance

components [12]. Variance (σ^2) is the statistical term that is used to indicate variability.

The variance in observed scores can be subdivided into the variance in the objects under study, in our example the persons (σ^2_p), the variance in observers (the two different PTs) (σ^2_{pt}), and the interaction between persons and PTs. We will call this latter term the residual variance ($\sigma^2_{residual}$).* The variance in persons (σ^2_p) represents the variability between persons, and σ^2_{pt} represents the variance due to systematic differences between PT_A and PT_B. The measurement error [error variance (σ^2_{error})] consists of either $\sigma^2_{residual}$ or of ($\sigma^2_{pt} + \sigma^2_{residual}$), depending on whether or not one wants to take into account systematic differences between the measurements (in our example PTs A and B). These systematic differences are usually considered to be part of the measurement error, because in practice, the measurements are performed by different PTs, and one is interested in the real values of the differences between the repeated measurements. However, if one is only interested in the ranking of patients, the systematic differences between the PTs are not important, and the error variance contains only $\sigma^2_{residual}$ [12].

The ICCs, which relate the measurement error to the variability between persons, are represented by the following formulas, for $ICC_{agreement}$ and $ICC_{consistency}$ [11], respectively:

$$ICC_{agreement} = \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_{pt} + \sigma^2_{residual}}$$

$$ICC_{consistency} = \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_{residual}}$$

We realize that these specifications ''agreement'' and ''consistency'' for the type of ICC is highly confusing, because both terms have others meanings in the field of reproducibility. For example, consistency is sometimes used as synonym of reproducibility. As this terminology of ICC types is used in general handbooks and landmark papers [11,12], we do not want to deviate from it. $ICC_{agreement}$ has the extra term σ^2_{pt} in the denominator to take the systematic difference between the PTs into account; the $ICC_{consistency}$ ignores systematic differences. Both types of ICCs are dependent on the heterogeneity of the population sample with respect to the characteristic of the study. We want to stress that both $ICC_{agreement}$ and $ICC_{consistency}$ are reliablity parameters (and not agreement parameters), although the term $ICC_{agreement}$ suggests otherwise.

The measurement error is represented by the standard error of measurement (SEM) and equals the square root of the error variance: $SEM = \sqrt{\sigma^2_{error}}$. This means that $SEM_{agreement} = \sqrt{(\sigma^2_{pt} + \sigma^2_{residual})}$ and $SEM_{consistency} = \sqrt{\sigma^2_{residual}}$. The SEM is a suitable parameter of agreement.

6. Illustration of ICC and SEM calculations in the example

Table 2 presents the values of the variance components for the affected and nonaffected shoulder. The variance components are estimated with SPSS (version 10.1), with the range of motion values as independent variable and persons and PTs as random factors, using the restricted maximum likelihood method. From these variance components, the above-mentioned SEMs can be calculated. For the affected shoulder:

$$SEM_{agreement\_AB} = \sqrt{(\sigma^2_{pt\_AB} + \sigma^2_{residual})} = \sqrt{(0 + 49.98)} = 7.07^\circ$$

$$SEM_{consistency\_AB} = \sqrt{\sigma^2_{residual}} = \sqrt{49.98} = 7.07^\circ$$

The $ICC_{agreement\_AB}$ for the affected shoulder can be calculated as follows:

$$ICC_{agreement\_AB} = \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_{pt\_AB} + \sigma^2_{residual}} = \frac{236.93}{236.93 + 0.00 + 49.98} = 0.83$$

And for the nonaffected shoulder:

$$ICC_{agreement\_AB} = \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_{pt\_AB} + \sigma^2_{residual}} = \frac{18.08}{18.08 + 0.11 + 45.91} = 0.28$$

In this example, $ICC_{agreement}$ and $ICC_{consistency}$ have roughly the same value, as the systematic differences between PT_A and PT_B were small, $\sigma^2_{pt\_AB}$ is almost 0. Therefore, we introduce a hypothetic physical therapist C (PT_C), who scores the range of motion of every patient 5° lower than PT_B. We added the (hypothetic) scores of PT_C to our dataset and recalculated the variance components (Table 2).

Table 2
Values of the variance components for the affected and nonaffected shoulder

Variance components	Affected shoulder	Nonaffected shoulder
σ^2_p	236.93	18.08
$\sigma^2_{pt\_AB}$	0.00	0.11
$\sigma^2_{pt\_AC}$	16.50	17.09
$\sigma^2_{residual}$	49.98	45.91

$\sigma^2_{pt\_AB}$ represents error due to systematic differences between PT_A and PT_B.

$\sigma^2_{pt\_AC}$ represents error due to systematic differences between PT_A and PT_C.

* $\sigma^2_{residual}$ is sometimes expressed as σ^2_{p*pt}, and represents the interaction between PTs and persons. As explained in an earlier paragraph, this variance omponent cannot be disentangled.

From now on we proceed with the example of PT_A and PT_C, because this will better illustrate the influence of systematic differences on the parameters for agreement and reliability. We will present only the calculations for the affected shoulder in the text. The results for both shoulders are presented in Table 3.

$$ICC_{agreement\_AC} = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_{pt\_AC}^2 + \sigma_{residual}^2} = \frac{236.93}{6.93 + 16.50 + 49.98} = 0.78$$

$$ICC_{consistency\_AC} = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_{residual}^2} = \frac{236.93}{236.93 + 49.98} = 0.83$$

Note that $ICC_{consistency}$ is not influenced by the systematic differences, but $ICC_{agreement\_AC}$ becomes smaller, because for this parameter systematic differences between PT_A and PT_C ($\sigma_{pt\_AC}^2$) are included in the measurement error.

7. Three ways to obtain SEM values

To facilitate and encourage the use of agreement parameters we will demonstrate how agreement parameters can be derived from the ICC formula, or can be calculated in other ways.

1. SEM values can easily be derived from the ICC formula, if all variance components are presented. In that case, the reader can calculate the ICC of his/her own choice. SEM is calculated as $\sqrt{\sigma_{error}^2}$, which equals $\sqrt{(\sigma_{pt}^2 + \sigma_{residual}^2)}$, if one wishes to take the systematic differences between the PTs into account, otherwise, it equals $\sqrt{\sigma_{residual}^2}$.

Table 3
Reproducibility of measurement of PT_A and PT_C

Parameters	Affected shoulder	Nonaffected shoulder
PT_A: Mean (SD)	69.49° (17.60°)	79.78° (7.60°)
PT_C: Mean (SD)	63.69° (16.25°)	73.88° (8.38°)
$PT_A - PT_C$: $Mean_{diff\_AC}$ (SD_{diff_AC})	5.80° (10.0°)	5.90° (9.58°)
$ICC_{agreement\_AC}$	0.78	0.22
$ICC_{consistency\_AC}$	0.83	0.28
$SEM_{agreement\_AC}$	8.15°	7.94°
$SEM_{consistency\_AC}$	7.07°	6.78°
Limits of $Agreement_{\_AC}$	−13.80°–25.40°	−12.88° −24.68°

Abbreviations: ICC: intraclass correlation coefficient; $Mean_{diff\_AC}$: Mean of the differences between PT_A and PT_B; PT_A: physical therapist A; PT_B: physical therapist B; SD: standard deviation; SD_{diff_AC}: standard deviation of the differences between PT_A and PT_B.

$$SEM_{agreement\_AC} = \sqrt{\sigma_{error}^2} = \sqrt{(\sigma_{pt\_AC}^2 + \sigma_{residual}^2)} = \sqrt{(16.50 + 49.98)} = 8.15°$$

$$SEM_{consistency\_AC} = \sqrt{\sigma_{error}^2} = \sqrt{\sigma_{residual}^2} = \sqrt{49.98} = 7.07°$$

2. The ICC formula can be transformed to $SEM = \sigma\sqrt{(1-ICC)}$ [1], in which σ represents the total variance (i.e., the denominator of the reliability formula). Only $SEM_{consistency}$ can be reproduced this way, by imputing the pooled SD of the first and second assessment for σ, and using $ICC_{consistency}$:

$$SEM_{consistency\_AC} = \sigma\sqrt{(1-ICC_{consistency\_AC})} = 16.94 * \sqrt{(1-0.826)} = 7.07°$$

Note that $SEM_{agreement\_AC}$ cannot be obtained in this way as calculated, because systematic errors are not represented in the pooled SD. Using $ICC_{agreement}$ and the pooled SD of the first and second assessment for σ would yield:

$$SEM_{agreement\_AC} = \sigma\sqrt{(1-ICC_{agreement\_AC})} = 16.94 * \sqrt{(1-0.781)} = 7.93° \neq 8.15°$$

The formula $SEM = \sigma\sqrt{(1-ICC)}$ is often used if information on the individual variance components is lacking. The ICC calculated in one study is then applied to a population sample for which the standard deviation (the total variance) is known. In this way, only a raw indication of the SEM can be obtained, because the ICC is heavily dependent on the heterogeneity of the characteristic under study in the population sample, and is thus, in theory, only applicable for a population with a similar heterogeneity.

3. The value of SEM can also be derived by dividing the SD of the mean differences between two measurements (SD_{diff}) by $\sqrt{2}$. The factor $\sqrt{2}$ is included because it concerns the difference between two measurements and errors occur in both measurements. Note that the SEM obtained by this formula is again $SEM_{consistency}$, because systematic error is not included in the SD. Thus, $SEM_{agreement\_AC}$ cannot be calculated in this way either.

$$SEM_{consistency\_AC} = SD_{diff\_AC}/\sqrt{2} = 10.00/\sqrt{2} = 7.07°$$

$SEM_{consistency\_AB}$ gives the same result, because PT_B and PT_C only differed by a systematic value.

8. Typical parameters for agreement and reliability

For repeated measurements on a continuous scale, as in our example, an ICC is the most appropriate reliability

parameter. An extensive overview of the various ICC formulas is provided by McGraw and Wong [11].

In our example, agreement was expressed as the percentage of observations lying between predefined values (Table 1). Presentation in this way makes sense in clinical practice, because every PT knows what 5° and 10° means. This measure was chosen because it can easily be interpreted by PTs [3]. However, the SEM is usually the basic parameter of agreement for measurements on a continuous scale. A method proposed by Bland and Altman [4], which assesses the limits of agreement is frequently used. These limits of agreement can be directly derived from the $SD_{diff} = (\sqrt{2} * SEM_{consistency})$.

9. Clinical interpretation

Agreement parameters are expressed on the actual scale of measurement, and not as reliability parameters as a dimensionless value between 0 and 1. This is an important advantage for clinical interpretation. If weights are measured in kilograms, the dimension of the SEM is kilograms. For example, if we know that a weighing scale has a SEM of 300 g, we know that we can use it to monitor adult body weight because changes of less than 1 kilogram are not important. The smallest detectable change (SDC) is based on this measurement error, and is defined as $1.96 * \sqrt{2} * SEM$.* With an SEM of 300 g, SDC is $1.96 * \sqrt{2} * 300g = 832g$. Obviously, one cannot use this scale to weigh babies or to weigh flour in the kitchen, because in these instances changes of less than 800 g are very important. The measurement error alone provides useful information when there is a clear conception of the differences that are important.

The situation is different in the case of unfamiliarity with scores. For example, if a new multiitem questionnaire is used to measure functional status on a scale from 0 to 50, an orthopaedic surgeon may want to know what a value of 14 points and an SEM of 2 points means, because she has no idea how many points of change represent clinically relevant change. By presenting an ICC she will know whether the instrument is able to discriminate between patients in the sample, but she still does not know whether the instrument is suitable for monitoring the functional status of her patients over time. This requires more information about the interpretation of scores. By assessing the scores of groups of mildly, moderately, and severely disabled patients a feeling for the meaning of scores will arise. Comparisons with other instruments will provide further insight into the meaning of values on the new measurement instrument. The assessments of minimally important changes in various measurements will also contribute to insight with regard to which (changes in) scores are clinically relevant [13,14]. Only this information makes it possible to assess whether the agreement parameter of a measurement instrument is sufficient to detect clinically relevant changes.

10. Conclusion

In this article we have shown the important difference between the parameters of reliability and agreement and their relationship. Agreement parameters will be more stable over different population samples than reliability parameters, as we observed in our shoulder example, in which the SEM was quite similar for the affected and the nonaffected shoulder. Reliability parameters are highly dependent on the variation in the population sample, and are only generalizable to samples with a similar variation. Reliability is clearly a characteristic of the performance of an instrument in a certain population sample. Agreement is more a characteristic of the measurement instrument itself. Agreement parameters are preferable in all situations in which the instrument will be used for evaluation purposes, which is often the case in medical research. Researchers and readers should be eager to apply and interpret the parameters of agreement and reliability correctly.

* The term 'smallest' detectable difference (SDD) is also used for this purpose.

References

[1] Streiner DL, Norman GR. Health Measurement Scales. A practical guide to their development and use. 3rd ed. New York: Oxford University Press Inc.; 2003.
[2] McDowell I, Newell C. Measuring health. A guide to rating scales and questionnaires. 2nd ed. New York: Oxford University Press Inc.; 1996.
[3] De Winter AF, Heemskerk MAMB, Terwee CB, Jans MP, Van Schaardenburg D, Scholten RJPM, Bouter LM. Inter-observer reproducibility of range of motion in patients with shoulder pain using a digital inclinometer. BMC Musculoskel Disord 2004;5:18.
[4] Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurements. Lancet 1986;i: 307–10.
[5] Nunnally JC, Bernstein IH. Psychometric theory. 3rd ed. New York: McGraw-Hill Inc.; 1994.
[6] Stratford PW, Goldsmith CH. Use of the standard error as a reliability index of interest: an applied example using elbow flexor strength data. Phys Ther 1997;77:745–50.
[7] De Vet HCW. Observer reliability and agreement. In: Armitage P, Colton T, editors, Encyclopedia biostatistica, Vol 4. Chichester: John Wiley & Sons, Ltd.; 1998. p. 3123–8.
[8] Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. J Chronic Dis 1987;40: 171–8.
[9] Bot SD, Terwee CB, Van der Windt DA, Bouter LM, Dekker J, De Vet HC. Clinimetric evaluation of shoulder disability questionnaires: a systematic review of the literature. Ann Rheum Dis 2004;63: 335–41.
[10] De Boer MR, Moll AC, De Vet HC, Terwee CB, Volker-Dieben HJ, Van Rens GH. Psychometric properties of vision-related quality of

life questionnaires: a systematic review. Ophthal Physiol Opt 2004;24:257–73.
[11] Shavelson RJ, Webb NM. Generalizability theory. A primer. London: Sage Publications; 1991.
[12] McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychol Methods 1996;1:30–46.
[13] Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. J Clin Epidemiol 2003;56: 395–407.
[14] Testa MA. Interpretation of quality-of-life outcomes. Issues that affect magnitude and meaning. Med Care 2000;38:II-166–74.